Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Discrete Algorithms 2 (2004) 271–287

JOURNAL OF DISCRETE ALGORITHMS

www.elsevier.com/locate/jda

Parametric multiple sequence alignment and phylogeny construction

David Fernández-Baca [a,*,1], Timo Seppäläinen [b,2], Giora Slutzki [a]

[a] *Department of Computer Science, Iowa State University, Ames, IA 50011, USA*
[b] *Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706, USA*

Abstract

Bounds are given on the size of the parameter-space decomposition induced by multiple sequence alignment problems where phylogenetic information may be given or inferred. It is shown that many of the usual formulations of these problems fall within the same integer parametric framework, implying that the number of distinct optima obtained as the parameters are varied across their ranges is polynomially bounded in the length and number of sequences.

Keywords: Alignment; Computational biology; Evolutionary tree; Multiple alignment; Phylogenetic tree; Sensitivity analysis

1. Introduction

Aligning two or more sequences to highlight their similarities or differences is among the most basic problems in computational biology. In a way, multiple (that is, three or more) sequence comparison is just a generalization of pairwise sequence comparison, a question that has been studied extensively (see [19]). However, from a practical point of view, multiple sequence comparison can be significantly more informative than pairwise comparison. This is because highly dissimilar DNA sequences may have similar functions. By considering many sequences jointly, one can exhibit evolutionary patterns that may

* Corresponding author.
E-mail addresses: fernande@cs.iastate.edu (D. Fernández-Baca), seppalai@math.wisc.edu (T. Seppäläinen), slutzki@cs.iastate.edu (G. Slutzki).

[1] Part of this work was conducted while the author was on leave at the University of California, Davis. Supported in part by the National Science Foundation under grants CCR-9520946 and CCR-9988348.
[2] Supported in part by the National Science Foundation under grant DMS-9801085.

1570-8667/$ – see front matter
doi:10.1016/S1570-8667(03)00078-9

be missed by pairwise comparison. On the other hand, when three or more sequences are involved, issues arise that are absent in pairwise alignment, making definitions more complex and the associated problems harder to solve efficiently [35]. One of these new facets is the role of evolutionary relationships between sequences. One possibility is to disregard these relationships, at least to a certain extent, by comparing all sequences against each other—the *sum-of-pairs* approach [5] is one example. Using or attempting to infer evolutionary relationships leads to a host of new problems. In one problem, called *phylogenetic alignment*, the input is a tree whose leaves are labeled by sequences and the objective is to find a labeling of the internal nodes that minimizes the total length of the tree, which is the sum of the (evolutionary) distances between adjacent sequences in the tree [32]. In another problem, *generalized phylogenetic alignment*, the input is a set $\mathcal{S}$ of sequences and one must find a sequence-labeled tree of minimum length wherein the elements of $\mathcal{S}$ are precisely the labels of the leaves of the tree [26]. Sum-of-pairs multiple alignment, phylogenetic alignment, generalized phylogenetic alignment, and some of their variants are known to be NP-hard [25,26]. The first two can be solved in time polynomial in the lengths of the sequences and exponential in their number. Implementations of several of these methods, often relying on heuristics, are available [11,15,29].

The optimum solutions to alignment problems depends on the various parameters used to compute inter-sequence distance or similarity—e.g., the weights of mismatches and spaces. While parameter choice can have a dramatic effect on alignment quality, there are no precise selection rules to rely on, and there is probably no single choice that is appropriate for all circumstances [36]. One approach to overcoming this difficulty is to examine the space of *all* parameter choices by conducting a *parametric analysis* [16]. This question has been studied for pairwise sequence comparison [13,20,22,24,36,37], to a lesser extent for sum-of-pairs multiple alignment [13,34], and hardly at all for phylogenetic alignment—see, however, [38]. Here we explore parametric multiple alignment and phylogenetic alignment.

One objective of parametric analysis is to establish upper bounds on the number of distinct *optimality regions*, i.e., maximal connected regions of the parameter space such that, within each region a single solution is optimal. Gusfield et al. [20], obtained the first results, proving, among other things, a $O(n^{2/3})$ upper bound on the number of regions for two-sequence global alignment, the shorter of which has length n. This was extended to obtain a $O(n^{2/3}k^{4/3})$ upper bound for the parametric sum-of-pairs alignment of k sequences of length n [13]. In the present paper, we argue that many multiple alignment schemes fall within the same "integer parametric" framework. This leads to the somewhat surprising result that the number of optimality regions for these problems is polynomial in both the lengths of the sequences and their number, even if the scoring is alphabet-dependent. Our bounds are consequences of the following observation: While the number of potential phylogenies and sequences labeling them is exponentially large, any scoring system based on affine functions whose coefficients are themselves functions of discrete features of alignments (e.g., number of mismatches, spaces, etc.) only allows a polynomially-bounded number of distinct cost functions to be optimal. The techniques used are uniform and straightforward; once the problems are formulated properly, a common structure emerges. Better bounds might be obtainable by a tighter analysis within our framework; however,

we suspect that significant advances will require deeper understanding of the combinatorial structure of the individual problems.

Main results The alignment problems studied here are classified according to whether the scoring is (i) local or global, (ii) distance-based or similarity-based, (iii) alphabet-dependent or alphabet-independent, (iv) dependent on the number of gaps or not. The input consists of k sequences, which, for simplicity, are assumed to have the same length n. Our results include the following bounds on the number of optimality regions.

1. A $O(n^{2/3}p^{2/3})$ bound for global multiple alignment under sum-of-pairs alphabet-independent similarity and distance scoring, with zero gap penalty, when the induced alignment of only p of the $\binom{k}{2}$ possible pairs is considered for computing the total score.
2. A $O(n^{5/3}p^{5/3})$ bound for the previous problem when the gap penalty varies. For the pairwise case ($p = 1$), this improves on earlier $O(n^2)$ bound by Gusfield et al. A $O(n^{5/3}p^{5/3})$ bound also holds for the local case when the gap penalty is zero. For the pairwise case, this improves on a $O(n^2)$ bound by Gusfield et al. [20].
3. A $O(n^{2/3}k^{4/3})$ bound for phylogenetic and generalized phylogenetic alignment under distance-based global alphabet-independent scoring with zero gap penalty. The bound goes up to $O(n^{5/3}k^{10/3})$ when the gap penalty is variable.
4. A $O(n^{2/3}k^{2/3})$ bound for star alignment (a special case of tree alignment) under global alphabet-independent distance-based scoring with zero gap penalty. This increases to $O(n^{5/3}k^{5/3})$ when the gap penalty is allowed to vary.
5. Polynomial bounds for sum-of-pairs, phylogenetic alignment, and generalized phylogenetic alignment problems under alphabet-dependent, global or local, similarity or distance scoring.

We use these results to argue that for all the above problems the entire parameter space decomposition can be constructed by computing polynomially-many optimum alignments.

Organization of the paper The main problems studied here, as well as some of their properties, are defined in Section 2. Section 3 discusses parametric analysis in a general context and consists of two parts. First, we obtain an upper bound on the number of optimality regions for parametric problems satisfying certain integrality conditions. Second, we describe a general approach to generating parameter-space decompositions. Parametric multiple alignments are discussed in Section 4. Section 5 presents some conclusions and open problems.

2. Preliminaries

We now give formal definitions and prove some of the basic properties of the problems whose parametric versions we shall study. The first part of this section introduces distance and similarity measures based on pairwise alignments. These notions are the basis for the scoring schemes used in multiple sequence comparison, which are discussed in the second

part. In what follows, Σ will denote an alphabet that includes a special *space* character "-". All input strings are assumed to be over $\Sigma \setminus \{-\}$.

2.1. Pairwise alignments

An *alignment* between two strings S_1 and S_2 is a pair of equal-length strings $\mathcal{A} = (S'_1, S'_2)$ where S'_1 and S'_2 are obtained by inserting space characters into S_1 and S_2 respectively, so that there is no character position in which both S'_1 and S'_2 have spaces. A *match* is a position in which S'_1 and S'_2 have the same character. A *mismatch* is a position in which S'_1 and S'_2 have different characters, neither of which is a "-". An *indel* is a position in which one of S'_1 and S'_2 has a "-". A *gap* is a sequence of one or more consecutive spaces in S'_1 or S'_2. Collectively, we call the matches, mismatches, indels, and gaps the *features* of $\mathcal{A}$. These features are used to compute the *value* of $\mathcal{A}$ according to a certain *scoring scheme*. In *local* scoring schemes the goal is to locate highly similar substrings. In *global* schemes the entire input strings are taken into account.

Global alignment We first consider alphabet-independent scoring schemes. Let $w_{\mathcal{A}}$, $x_{\mathcal{A}}$, $y_{\mathcal{A}}$, and $z_{\mathcal{A}}$ denote, respectively, the number of matches, mismatches, indels, and gaps in an alignment $\mathcal{A}$, and let α, β, and γ be the mismatch, indel, and gap penalties, respectively. Penalties are assumed to be nonnegative.

The similarity value of an alignment $\mathcal{A}$ is given by

$$\sigma_{\mathcal{A}} = w_{\mathcal{A}} - \alpha x_{\mathcal{A}} - \beta y_{\mathcal{A}} - \gamma z_{\mathcal{A}}. \tag{1}$$

The global *similarity* between sequences S_1 and S_2 is defined as

$$\mathrm{sim}(S_1, S_2) = \max\{\sigma_{\mathcal{A}}\colon\ \mathcal{A} \text{ is an alignment of } S_1 \text{ and } S_2\}. \tag{2}$$

The distance value of an alignment $\mathcal{A}$ is

$$\delta_{\mathcal{A}} = \alpha x_{\mathcal{A}} + \beta y_{\mathcal{A}} + \gamma z_{\mathcal{A}}. \tag{3}$$

The *distance* between S_1 and S_2 is

$$\mathrm{dist}(S_1, S_2) = \min\{\delta_{\mathcal{A}}\colon\ \mathcal{A} \text{ is an alignment of } S_1 \text{ and } S_2\}. \tag{4}$$

Note that the quantities defined in Eqs. (1)–(4) are all functions of the penalties α, β, and γ. Schemes (1) and (3) are related by the lemma below, in which the mismatch, indel, and gap penalties are given by triples (α, β, γ).

Lemma 1. *Under global alphabet-independent scoring,*

$$\sigma_{\mathcal{A}}(\alpha, \beta, \gamma) = \frac{n+m}{2} - \delta_{\mathcal{A}}(\alpha + 1, \beta + 1/2, \gamma),$$

where n and m are the lengths of the input strings. Therefore, a pairwise alignment has maximum similarity score at (α, β, γ) if and only if it has minimum distance score at $(\alpha + 1, \beta + 1/2, \gamma)$.

Proof. Since every pairwise alignment $\mathcal{A}$ satisfies $2w_{\mathcal{A}} + 2x_{\mathcal{A}} + y_{\mathcal{A}} = n + m$ [20], the similarity score of $\mathcal{A}$ (1) can be re-expressed as

$$\begin{aligned}\sigma_{\mathcal{A}}(\alpha, \beta, \gamma) &= \frac{n+m}{2} - (\alpha+1)x_{\mathcal{A}} - \left(\beta + \frac{1}{2}\right)y_{\mathcal{A}} - \gamma z_{\mathcal{A}} \\ &= \frac{n+m}{2} - \delta_{\mathcal{A}}(\alpha+1, \beta+1/2, \gamma),\end{aligned}$$

where the second line follows from the definition of distance score (3). The rest of the lemma follows immediately. □

Note that we can assume without loss of generality that the mismatch penalty α in (3) is one, since changing its value only affects the magnitude, but not the relative values, of the alignments. Thus, there are effectively only two parameters to be chosen for distance scoring. By Lemma 1, this is also true for similarity scoring.

Alphabet-dependent scoring schemes depend on a symmetric $|\Sigma| \times |\Sigma|$ *substitution matrix* α, where $\alpha(s, t)$ is the cost of lining up character s with character t, which may be positive, zero, or negative (the latter even for $s = t$). Widely-used families of matrices for protein alignment are PAM [6] and BLOSUM [23]. Recall that we assume that Σ contains "-"; $\alpha(t, \text{-})$ is thus the indel penalty. The similarity score of an alignment $\mathcal{A}$ is now given by

$$\sigma_{\mathcal{A}} = -\gamma z_{\mathcal{A}} + \sum_{\{s,t\} \subseteq \Sigma} \alpha(s, t) \cdot x_{\mathcal{A}}(s, t), \tag{5}$$

where $x_{\mathcal{A}}(s, t)$ is the number of times character s is lined up with character t in $\mathcal{A}$ and $z_{\mathcal{A}}$ is the number of gaps in $\mathcal{A}$. The similarity between two sequences is obtained by applying this scoring scheme in (2).

The alphabet-dependent distance score of $\mathcal{A}$, $\delta_{\mathcal{A}}$, can be defined in the same way as (5), except that the "$-\gamma z_{\mathcal{A}}$" term is replaced by "$+\gamma z_{\mathcal{A}}$". For both similarity and distance, the total number of parameters is $(|\Sigma|^2 + |\Sigma|)/2 + 1$: the number of entries in the substitution matrix plus the gap penalty.

Local alignment For two strings S and R, we write $S \sqsubseteq R$ if S is a substring of R. The *local* similarity between S_1 and S_2, denoted $\text{sim}_L(S_1, S_2)$, is defined as

$$\text{sim}_L(S_1, S_2) = \max\{\sigma_{\mathcal{A}}\colon \mathcal{A} \text{ is an alignment of } S_1' \sqsubseteq S_1 \text{ with } S_2' \sqsubseteq S_2\}. \tag{6}$$

The scoring scheme used in the definition above may be alphabet-dependent or -independent. Note that the global similarity between two strings is a lower bound on their local similarity. Note also that while it is straightforward to define local distance measures using minimization instead of maximization, it makes no sense to do so under alphabet-independent scoring (3), since one can trivially achieve an optimum score of zero by aligning empty substrings from each of S_1 and S_2 (this assumes that $\alpha, \beta, \delta > 0$). Thus, in the local case, the alphabet-independent versions of distance and similarity are *not* related by Lemma 1, which means that even though global similarity effectively depends on two parameters, local similarity still depends on three. On the other hand, alphabet-dependent local distance is a valid measure, which, like local similarity, depends on $(|\Sigma|^2 + |\Sigma|)/2 + 1$ parameters.

Bounds on the features of alignments We will need the following facts about pairwise alignment, which were proved in [20]. As before $w_{\mathcal{A}}$, $x_{\mathcal{A}}$, $y_{\mathcal{A}}$, and $z_{\mathcal{A}}$ denote the number of matches, mismatches, indels, and gaps in an alignment $\mathcal{A}$. Let n and m denote the lengths of the input strings, where $n \leqslant m$.

Lemma 2. *For any pairwise global or local alignment $\mathcal{A}$, $w_{\mathcal{A}} + x_{\mathcal{A}} \leqslant n$.*

Lemma 3. *For any global or local alignment $\mathcal{A}$, $z_{\mathcal{A}} \leqslant y_{\mathcal{A}} \leqslant m + n$. Moreover, if $\mathcal{A}$ is global, $y_{\mathcal{A}} \geqslant m - n$.*

2.2. Multiple alignments

A *multiple alignment* $\mathcal{A}$ of strings $S_1, \ldots, S_k$, where S_i has length n_i, is obtained by inserting spaces in each string to obtain strings of the same length l. The result is a matrix with k rows and l columns, such that each character and space of each string appears in exactly one column. $\mathcal{A}$ induces a pairwise alignment of S_i and S_j in a natural way: remove all rows of $\mathcal{A}$ except those corresponding to S_i and S_j and strike out any columns containing two spaces. This will be called the *induced pairwise alignment* of S_i and S_j.

The following generalization of two-sequence alignment was considered in [3,4]; it is used in the MSA package for multiple sequence alignment [15,29].

Weighted sum-of-pairs alignment (similarity version).

Input: A set of sequences $\mathcal{S} = \{S_1, \ldots, S_k\}$ and a $k \times k$ matrix $B = [b_{ij}]$.
Question: Find a multiple alignment $\mathcal{A}$ for $\mathcal{S}$ maximizing $\sum_{i<j} b_{ij} \boldsymbol{\sigma}_{\mathcal{A}(i,j)}$, where $\boldsymbol{\sigma}_{\mathcal{A}(i,j)}$ is the similarity score of the pairwise alignment between S_i and S_j induced by $\mathcal{A}$.

The scoring scheme can be global or local, alphabet-dependent or alphabet-independent. Distance versions of this problem can be defined in the obvious way, using minimization instead of maximization and appropriate scoring schemes. Sum of pairs alignment is known to be NP-hard [25] (see also [27,28]).

We note that, for practical reasons, the definition of a gap in a multiple alignment does not always correspond to a gap in one of the induced pairwise alignments. This fact will not impact the parametric analysis of Section 4 significantly. We refer the reader to [2,3,15] for further discussion on gap scoring.

The following result is an analog to Lemma 1.

Lemma 4. *Under global alphabet-independent weighted sum-of-pairs scoring, a multiple alignment has maximum similarity score at (α, β, γ) if and only if it has minimum distance score at $(\alpha + 1, \beta + 1/2, \gamma)$.*

Proof. By Lemma 1, the similarity and distance scores of the induced pairwise alignment between S_i and S_j, denoted $\boldsymbol{\sigma}_{\mathcal{A}(i,j)}$ and $\boldsymbol{\delta}_{\mathcal{A}(i,j)}$, respectively, are related by

$$\boldsymbol{\sigma}_{\mathcal{A}(i,j)}(\alpha, \beta, \gamma) = n - \boldsymbol{\delta}_{\mathcal{A}(i,j)}(\alpha + 1, \beta + 1/2, \gamma).$$

Thus,

$$\sum_{i<j} b_{ij}\sigma_{\mathcal{A}(i,j)}(\alpha,\beta,\gamma) = n\sum_{i<j} b_{ij} - \sum_{i<j} b_{ij}\delta_{\mathcal{A}(i,j)}(\alpha+1,\beta+1/2,\gamma).$$

Since $n\sum_{i<j} b_{ij}$ is fixed, it follows that the similarity score of $\mathcal{A}$ is maximized at (α,β,γ) if and only if its distance score is minimized at $(\alpha+1,\beta+1/2,\gamma)$. □

Thus, for the global alphabet-independent case, the score is a function of only two parameters. For the local alphabet-independent case, the score is a function of two parameters for distance measures and three for similarity measures (since the mismatch penalty must be considered, in addition to the indel and gap penalties). For the alphabet-dependent case, the score is still a function of $(|\Sigma|^2+|\Sigma|)/2+1$ parameters.

In the next two families of problems, evolutionary history is used and/or inferred. We define them for distance measures; similarity versions can be defined in the obvious way. Alphabet-dependent or -independent scoring can be used.

We need some definitions. A *phylogeny* for a set of sequences $\mathcal{S}$ is a tree T with $|\mathcal{S}|$ leaves, where every internal node has degree at least three and each element of $\mathcal{S}$ labels a distinct leaf of T. An *internal labeling* for T is an assignment of sequences over $\Sigma\setminus\{-\}$ to the internal nodes of T. The *length* of an internally-labeled phylogeny T is the sum of the pairwise distances between the labels of adjacent nodes.

Phylogenetic alignment.

Input: A phylogeny T for a set of sequences $\mathcal{S}$.
Question: Find an internal labeling for T that minimizes the total length of the resulting tree.

While this problem is NP-hard [25], it can be solved in polynomial time if the number of sequences is fixed [31–33]. An important special case is *star alignment*, where the tree T has only one internal node [5]; this problem is also NP-hard [25].

Given a solution to the phylogenetic alignment problem, one can derive a multiple alignment $\mathcal{A}$ for the sequences labeling the phylogeny that is *consistent* with it in the following sense: The value of the induced pairwise alignment for any two sequences labeling neighbors in the tree equals the distance between the sequences [19]. One can obtain a multiple alignment for $\mathcal{S}$ by striking out the rows of $\mathcal{A}$ that do not correspond to elements of $\mathcal{S}$. However, the labels on the internal nodes can be valuable, as they represent hypothetical ancestors to the elements of $\mathcal{S}$.

The next problem, which is NP-hard [26], is related to phylogenetic alignment, but the tree itself is not given.

Generalized phylogenetic alignment.

Input: A set of sequences $\mathcal{S}$.
Question: Find a minimum-length internally-labeled phylogeny for $\mathcal{S}$.

Phylogenetic alignment and its generalized version have similarity variants where the goal is to find a solution that maximizes the similarity score.

To prevent spurious matches between internal nodes, we make the following assumption: there exists a multiple alignment $\mathcal{A}$ consistent with the labels of T, each of whose columns contains at least one character from one of the sequences in $\mathcal{S}$.

One distinction between phylogenetic and SP alignment is that for the former similarity and distance measures are *not* equivalent in either the local or global cases. This equivalence for the SP problem is in part a consequence of knowing in advance the lengths of the strings involved in the pairwise comparisons. This is not true for phylogeny problems. Thus, for global and local alphabet-independent phylogenetic alignment, the score is a function of two parameters for distance measures and three for similarity measures. For the alphabet-dependent case, the score is a function of $(|\Sigma|^2 + |\Sigma|)/2 + 1$ parameters. We can, however, still establish useful bounds on the number of features.

Lemma 5. *For any alignment $\mathcal{A}$ of a set $\mathcal{S}$ of k sequences of length n to a phylogeny with r internal nodes, $w_{\mathcal{A}} + x_{\mathcal{A}} \leqslant nkr$ and $z_{\mathcal{A}} \leqslant y_{\mathcal{A}} \leqslant nk(k + 2r - 1)$.*

Proof. Let T be the input phylogeny for $\mathcal{S}$. Each unlabeled node in T is assigned a sequence of length at most nk, since each of its characters must line up with a character in some sequence in $\mathcal{S}$. By Lemma 2, the total number of matches and mismatches in the induced pairwise alignment between any two adjacent sequences in T is at most equal to the length of the shorter sequence. Thus, the contribution of an edge in T to the total number of matches and mismatches is n if one endpoint is an element of $\mathcal{S}$ and at most nk if neither endpoint is in $\mathcal{S}$. The total number of edges in the latter category is at most $r - 1$, while the number of edges in the former category is at most k. This establishes the bound for $w_{\mathcal{A}} + x_{\mathcal{A}}$.

To bound $y_{\mathcal{A}}$ and $z_{\mathcal{A}}$, we use Lemma 3, noting that edges where one endpoint is in $\mathcal{S}$ contribute at most $n + nk$ indels, while those where neither endpoint is in $\mathcal{S}$ contribute at most $2nk$ to the total. □

Improvements are possible on the previous bound for special cases. For global alignments, one can get a joint bound for $w_{\mathcal{A}}$, $x_{\mathcal{A}}$, and $y_{\mathcal{A}}$ of $w_{\mathcal{A}} + x_{\mathcal{A}} + y_{\mathcal{A}}/2 \leqslant nk(k + 2r - 1)/2$. Since this fact will not be used, we omit its proof. Another improvement can be obtained for optimal star alignments.

Lemma 6. *Let $\mathcal{A}$ be an optimal star alignment for a set of k sequences of length n under alphabet-independent distance-based global scoring. Then $y_{\mathcal{A}}, z_{\mathcal{A}} \leqslant nk$.*

Proof. We will argue that $y_{\mathcal{A}} \leqslant nk$, from which the bound on $z_{\mathcal{A}}$ follows.

Let X_c denote the sequence labeling the center of the star. If X_c is the empty string, the total distance is $\beta nk + \gamma k$; this value is an upper bound on the cost of the optimum alignment.

Now suppose $y_{\mathcal{A}} \geqslant nk + 1$. If $z_{\mathcal{A}} \geqslant k$, the score of $\mathcal{A}$ is greater than our upper bound and hence $\mathcal{A}$ cannot be optimum. Thus, suppose $z_{\mathcal{A}} < k$. If, for all i, the induced pairwise alignment between X_c and S_i has one or more indels, then $z_{\mathcal{A}} \geqslant k$. So assume $S_1, \ldots, S_l$

have no indels in their induced pairwise alignments with X_c. Then X_c must be of length n and the total number of indels in $\mathcal{A}$ is at most $2n(k-l)$. Thus, we must have $2n(k-l) \geqslant nk+1$, and hence $2(k-l) \geqslant k+1/n$. Now, $2(k-l)$ is also a lower bound on the number of gaps, since each one of sequences $S_{l+1}, \ldots, S_k$ contributes at least two indels to $\mathcal{A}$. Thus, if there are any indels in the induced pairwise alignment of X_c and S_i $(i > l)$, there must be at least two gaps. Therefore, $z_{\mathcal{A}} \geqslant k+1/n$, a contradiction. □

3. Parametric analysis

In this section, we consider two issues that arise in parametric analysis: finding the number of distinct optimal solutions attained as the parameters are varied across their range and generating all of these solutions. We study these questions within a framework that encompasses a broad class of problems. We first need some definitions.

Let $\lambda = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^d$ denote a parameter vector. An *affine parametric combinatorial optimization problem* is given by a finite set $X \subseteq \mathbb{R}^{d+1}$ of *feature vectors*, where each $\mathbf{x} \in X$ has an associated parametric cost function

$$f_{\mathbf{x}}(\lambda) = x_0 + \sum_{i=1}^{d} \lambda_i x_i. \tag{7}$$

The optimum solution at λ is given by

$$F(\lambda) = \min_{\mathbf{x} \in X} f_{\mathbf{x}}(\lambda). \tag{8}$$

We write "min" in the definition above for concreteness; the concepts and results to follow have analogs for maximization problems.

Since it is the lower envelope of a finite set of affine functions, F is piecewise affine. F induces a partition of the parameter space $\mathbb{R}^d$ into d-dimensional convex polyhedral *optimality regions*, such that $F(\lambda)$ is attained by a single function $f_{\mathbf{x}}$ for all λ in the interior of each such region. This subdivision of $\mathbb{R}^d$ is known as the *minimization diagram* of F [1].

Our framework encompasses all the alignment problems defined in Section 2 (in fact, it also covers problems that are unrelated to sequence alignment). For instance, in the pairwise alignment problem under global, alphabet-independent scoring, a feature vector is given by the number of matches, mismatches, indels, and gaps in some alignment and the parameters are the mismatch, indel, and gap penalties. Note that several feasible solutions (i.e., alignments) may have the same feature vector. Thus, even though a single *cost function* attains the optimal value for each region of F's minimization diagram, there might be several alignments with the same feature vector that are co-optimal within the region.

3.1. The number of optimality regions

We now prove upper bounds on the number of optimality regions for parametric problems of the form (8). We begin with the following result, which is implicit in the work of Gusfield et al. [20,21].

Lemma 7. *If $X \subseteq \{0, \ldots, N\}^2$ for some positive integer N, then $F(\lambda)$ induces $O(N^{2/3})$ optimality regions in $\mathbb{R}$.*

Proof. We rely on the following fact, which is shown in [21] (see also [13]):

(*) Let $\{a_i/b_i\}_{1 \leqslant i \leqslant k}$ be a set of (distinct) irreducible fractions with positive numerators and denominators such that $\sum_{i=1}^{k} a_i, \sum_{i=1}^{k} b_i \leqslant N$. Then $k = O(N^{2/3})$.

Let us denote a feature vector $\mathbf{x}$ by (x, y) and its cost by $f_{\mathbf{x}}(\lambda) = x + \lambda y$. Then, $F(\lambda)$ is a non-decreasing piecewise affine function consisting of a sequence of line segments. Hence, if (x_i, y_i) and (x_{i+1}, y_{i+1}) denote the intercept and slope of the ith and $(i+1)$st segments of F, then $x_i < x_{i+1}$ and $y_i > y_{i+1}$. The λ-value of the meeting point between the ith and $(i+1)$st segments of F is $\Delta x_i/\Delta y_i$, where $\Delta x_i = x_{i+1} - x_i$ and $\Delta y_i = y_i - y_{i+1}$. Thus, $\Delta x_1/\Delta y_1 < \Delta x_2/\Delta y_2 < \cdots < \Delta x_s/\Delta y_s$, where s is one less than the number of segments of F. Since the numerators and denominators of these fractions are nonnegative integers and $\sum_{i=1}^{s} \Delta x_i, \sum_{i=1}^{s} \Delta y_i \leqslant N$, (*) applies, implying that F has $O(N^{2/3})$ optimality regions. □

Lemma 8. *If $X \subseteq \{0, \ldots, N\}^{d+1}$ for some positive integer N, then $F(\lambda)$ induces $O(N^{d-1/3})$ optimality regions in $\mathbb{R}^d$.*

Proof. By induction on d. The basis, $d = 1$, follows from Lemma 7. For $d > 1$, define $X_j = \{\mathbf{x} \in X\colon x_d = j\}$ and $h_{\mathbf{x}}(\lambda) = x_0 + \sum_{i=1}^{d-1} \lambda_i x_i$. Then, we can express F as

$$F(\lambda) = \min_{j=0,\ldots,N} \Big(\big(\min_{\mathbf{x} \in X_j} h_{\mathbf{x}}(\lambda) \big) + \lambda_d j \Big).$$

By hypothesis, $F'_j(\lambda) = \min_{\mathbf{x} \in X_j} h_{\mathbf{x}}(\lambda)$ induces $O(N^{(d-1)-1/3})$ regions in $\mathbb{R}^{d-1}$. Thus, $g_j(\lambda) = F'_j(\lambda) + \lambda_d j$ induces a subdivision of $\mathbb{R}^d$ into $O(N^{(d-1)-1/3})$ cylindrically-shaped optimality regions whose boundary lines are parallel to the λ_d-axis. Since F is the lower envelope of $N+1$ such g_i's, it induces $O(N \cdot N^{(d-1)-1/3}) = O(N^{d-1/3})$ optimality regions in $\mathbb{R}^d$. □

The following observation generalizes a result in [20].

Lemma 9. *Suppose that $X \subseteq A_0 \times \cdots \times A_d$, where each A_i is a set of N_i distinct real values. Then, $F(\lambda)$ induces at most $2(\prod_{i=0}^{d} N_i)/\max_{0 \leqslant i \leqslant d} N_i$ optimality regions in $\mathbb{R}^d$.*

Proof. Assume first that $N_0 = \max_{0 \leqslant i \leqslant d} N_i$. Consider any $\mathbf{x}, \mathbf{y} \in X$ such that $F(\lambda') = f_{\mathbf{x}}(\lambda')$ and $F(\lambda'') = f_{\mathbf{y}}(\lambda'')$ for some λ', λ''. If $x_0 < y_0$, then we must have $x_i \neq y_i$ for some $i \in \{1, \ldots, d\}$, for otherwise $F(\lambda) < f_{\mathbf{y}}(\lambda)$ for all λ. Thus, out of all $(d+1)$-tuples $(x_0, \ldots, x_d)$ whose last d entries are equal, at most one is associated with a feasible solution that is optimal at some point. Hence, there are at most $\prod_{i=1}^{d} N_i$ functions that are optimal at some point, which also bounds the number of optimality regions of F.

If $N_j = \max_{0 \leqslant i \leqslant d} N_i$, $j \neq 0$, we simply divide through by λ_j, redefining the parameters, and repeat the above analysis. However, an extra factor of 2 appears, accounting

for the fact that this analysis actually has two similar cases, depending on whether λ_j is positive or negative. □

3.2. Constructing the minimization diagram

Algorithms for constructing the minimization diagram of parametric problems have been proposed before (see, e.g., [10,14,17]), mostly for the one- and two-parameter cases. Here we sketch an approach that appears to be part of the folklore,[3] but deserves to be more widely known.

An *evaluation* of F at $\lambda = (\lambda_1, \ldots, \lambda_d)$ consists of finding the feature vector $\mathbf{x} \in X$ such that $F(\lambda) = f_{\mathbf{x}}(\lambda)$ (note that we do not specify whether or not the actual optimum solution is returned). Evaluating $F(\lambda)$ is equivalent to computing the equation of the supporting hyper-plane of the convex set $B_F = \{(\lambda_1, \ldots, \lambda_d, z)\colon\ z \leqslant F(\lambda)\}$ at the point $(\lambda_1, \ldots, \lambda_d, F(\lambda))$. This operation has been called a *hyper-plane probe* by Dobkin et al. [7,8], who studied the problem of reconstructing a convex object from a sequence of such probes. Constructing B_F (or, equivalently, F) from repeated evaluations of F is one instance of this problem.

Theorem 10 (Dobkin et al. [7,8]). *F can be computed with $O(m + dv)$ evaluations, where m and v are, respectively, the number of optimality regions and vertices of the minimization diagram.*

For brevity, we omit the details of the probing algorithm, which is explained fully in [7,8]. The 1- and 2-parameter algorithms of [10] and [14] can be viewed as special cases. The probing process returns successive elements of a set H of half-spaces in $\mathbb{R}^d$ whose intersection equals B_F. Actually generating F is equivalent to computing a description of the boundary of this intersection. A good practical algorithm to do so is the beneath-beyond method [30], whose running time is $O(s|H|)$, where s is the size of the output (i.e., the total number of faces of dimension zero through d). Observe that $|H| = m$. Furthermore, it is a consequence of the Upper Bound Theorem [9] that if there are m optimality regions, $s = O(m^{\lceil d/2 \rceil})$. Thus, the minimization diagram of F can be constructed in time $O((m + dv)t_F + m^{\lceil d/2 \rceil + 1})$, where t_F is the time to evaluate F at any given λ.

4. Parametric multiple alignments

We now present upper bounds on the number of optimality regions for the multiple alignment problems of Section 2. To a certain extent, the results are independent of the kind of alignment problem we are dealing with, as long as we have bounds on the number of distinct values for the coefficients of the objective functions. The arguments are similar: We first show how the problem falls within the scope of Lemmas 7, 8 or 9 and then invoke the appropriate bound.

[3] Naoki Katoh, personal communication.

4.1. SP alignments

Our first results concern weighted sum-of-pairs (SP) alignments. We assume that the weight matrix is fixed. We first consider 0–1 weight matrices, a problem we refer to as 0–1 *SP alignment*. For the next three results, p denotes the number of non-zero entries above the main diagonal of the weight matrix $B = [b_{ij}]$. Observe that for pairwise alignment, $p = 1$. Given a multiple alignment $\mathcal{A}$, w_{ij}, x_{ij}, y_{ij}, and z_{ij} denote the number of matches, mismatches, indels, and gaps in the induced pairwise alignment for sequences i and j. By Lemmas 2 and 3, each of these values is at most $2n$. Thus,

$$0 \leqslant \sum_{1 \leqslant i < j \leqslant k} b_{ij} w_{ij}, \sum_{1 \leqslant i < j \leqslant k} b_{ij} x_{ij}, \sum_{1 \leqslant i < j \leqslant k} b_{ij} y_{ij}, \sum_{1 \leqslant i < j \leqslant k} b_{ij} z_{ij} \leqslant 2np. \tag{9}$$

We start with global alignment. Part (a) of the following result contains an earlier bound by Gusfield et al. [20] for pairwise alignments as a special case. On the other hand, when specialized to pairwise alignment, part (b) improves an earlier bound [20] by a factor of $n^{1/3}$.

Theorem 11. *The number of optimality regions for* alphabet-independent *parametric* 0–1 *SP alignment under* global *similarity or distance measures is*

(a) $O(n^{2/3}p^{2/3})$ *if the gap penalty is zero, and*
(b) $O(n^{5/3}p^{5/3})$ *if the gap penalty is variable.*

Proof. By Lemma 4 it suffices to consider distance measures. In this case, the distance value of a multiple alignment $\mathcal{A}$ is

$$\delta_{\mathcal{A}} = \sum_{1 \leqslant i < j \leqslant k} b_{ij} x_{ij} + \beta \sum_{1 \leqslant i < j \leqslant k} b_{ij} y_{ij} + \gamma \sum_{1 \leqslant i < j \leqslant k} b_{ij} z_{ij}. \tag{10}$$

Now, parts (a) and (b) follow by applying Lemma 8 for $d = 1$ and $d = 2$, respectively, with $N = 2np$, where the latter is valid by (9). □

The next theorem deals with local alignments. For the pairwise case, parts (a) and (b) improve on earlier bounds [20] by a factor of $n^{1/3}$.

Theorem 12. *The number of optimality regions for* alphabet-independent *parametric* 0–1 *SP alignment under* local *similarity measures is*

(a) $O(n^{5/3}p^{5/3})$ *if the gap penalty is zero, and*
(b) $O(n^{8/3}p^{8/3})$ *if the gap penalty varies.*

Proof. The total similarity score of a multiple alignment $\mathcal{A}$ is given by

$$\begin{aligned} \sigma_{\mathcal{A}} = & \sum_{1 \leqslant i < j \leqslant k} b_{ij} w_{ij} - \alpha \sum_{1 \leqslant i < j \leqslant k} b_{ij} x_{ij} \\ & - \beta \sum_{1 \leqslant i < j \leqslant k} b_{ij} y_{ij} - \gamma \sum_{1 \leqslant i < j \leqslant k} b_{ij} z_{ij}. \end{aligned} \tag{11}$$

Now, parts (a) and (b) follow from applying Lemma 8 for $d = 2$ and $d = 3$ with $N = 2np$. □

For alphabet-dependent scoring, we can prove a polynomial bound when the alphabet size is bounded.

Theorem 13. *The number of optimality regions for* alphabet-dependent *parametric global or local* 0–1 *SP alignment under distance and similarity measures is* $(np)^{O(|\Sigma|^2)}$.

Proof. By Eq. (5), the similarity score of an alignment $\mathcal{A}$ is

$$\begin{aligned}\sigma_{\mathcal{A}} &= -\gamma \sum_{1\leqslant i<j\leqslant k} b_{ij} z_{ij} + \sum_{1\leqslant i<j\leqslant k} \sum_{\{s,t\}\subseteq\Sigma} \alpha(s,t)\cdot x_{ij}(s,t) \\ &= -\gamma \sum_{1\leqslant i<j\leqslant k} b_{ij} z_{ij} + \sum_{\{s,t\}\subseteq\Sigma} \alpha(s,t) \sum_{1\leqslant i<j\leqslant k} b_{ij} x_{ij}(s,t), \end{aligned} \tag{12}$$

where z_{ij} and $x_{ij}(s,t)$ are, respectively, the number of gaps and the number of times character s is lined up with character t in the pairwise alignment between strings i and j induced by $\mathcal{A}$. The claim follows from Lemma 9, since $\sum_{1\leqslant i<j\leqslant k} b_{ij} x_{ij}(s,t)$ can take on $O(np)$ distinct values. The argument for distance scoring is completely analogous. □

We now consider the situation where the weights are arbitrary real values.

Theorem 14. *The number of optimality regions for global or local parametric SP alignment is* $n^{O(k^2)}$ *for the alphabet-independent case and* $n^{O(k^2|\Sigma|^2)}$ *for the alphabet-dependent case under similarity and distance measures.*

Proof. We consider only alphabet-dependent scoring; the other cases are similar. The cost of an alignment is then given by (12). By Lemma 2, each x_{ij} can take on $O(n)$ distinct values. Thus, since B is fixed, $\sum_{i<j} b_{ij} x_{ij}(s,t)$ can take on $n^{O(k^2)}$ distinct values. Because there are $O(|\Sigma|^2)$ parameters, the claim follows from Lemma 9. □

Finally, consider a two-parameter problem that *does not* involve a weight matrix. Here the goal is to analyze the trade-offs between match, mismatch, and indel penalties under alphabet-dependent scoring, when the substitution matrix is fixed. Given a multiple alignment $\mathcal{A}$, and $1 \leqslant i < j \leqslant k$, define three quantities,

$$M_{ij}(\mathcal{A}) = \sum_{t\in\Sigma\setminus\{-\}} \alpha(t,t) x_{ij}(t,t), \qquad MS_{ij}(\mathcal{A}) = \sum_{s,t\in\Sigma\setminus\{-\},s\neq t} \alpha(s,t) x_{ij}(s,t),$$

$$S_{ij}(\mathcal{A}) = \sum_{t\in\Sigma\setminus\{-\}} \alpha(t,-) x_{ij}(t,-).$$

The total score of $\mathcal{A}$ is the sum of the pairwise scores:

$$\sigma_{\mathcal{A}}(\lambda,\mu) = \sum_{i<j} M_{ij} - \lambda \sum_{i<j} MS_{ij} - \mu \sum_{i<j} S_{ij}. \tag{13}$$

We refer to the problem of finding a maximum-score alignment under the above scoring scheme as the *SP trade-off problem*. Gusfield et al. [20] considered the pairwise version of the problem, proving a sub-exponential bound on the number of optimality regions encountered in traversing the (λ, μ)-plane along any line. When the entries of the substitution matrix are small integers, as is often true for PAM and BLOSUM matrices used in practice, we can prove a better bound.

Theorem 15. *Suppose that for $s, t \in \Sigma$, $\alpha(s,t) \in \mathbb{Z}$ and $|\alpha(s,t)| \leqslant U$, $U \in \mathbb{Z}$. Then, the total number of optimality regions induced by the SP trade-off problem on the (λ, μ) plane is $O(n^{5/3}U^{5/3}k^{10/3})$.*

Proof. It follows from Lemma 8 since $\sum_{i<j} M_{ij}$, $\sum_{i<j} MS_{ij}$, and $\sum_{i<j} S_{ij}$ are $O(nUk^2)$. □

4.2. Phylogenetic alignments

Abusing terminology, we shall call a feasible solution to the phylogenetic alignment problem a *phylogenetic alignment* (or, simply, an alignment). An alignment will be viewed as consisting of both an internal labeling for the input phylogeny T and, for each edge of T a pairwise alignment between the sequences labeling its endpoints. For the generalized case, a feasible solution (also called an *alignment*) will consist of a phylogeny together with a phylogenetic alignment.

In the following theorems, $\mathcal{S}$ denotes the set of sequences, k the size of this set, and n the length of each sequence. We first study alphabet-independent scoring. Our bounds are the same for phylogenetic and generalized phylogenetic alignment.

Theorem 16. *The number of optimality regions for parametric phylogenetic and generalized phylogenetic alignment under* alphabet-independent *scoring is*

(a) $O(n^{2/3}k^{4/3})$ *under the distance measure if the gap penalty is zero,*
(b) $O(n^{5/3}k^{10/3})$ *under the distance measure if the gap penalty is allowed to vary,*
(c) $O(n^{5/3}k^{10/3})$ *under the similarity measure if the gap penalty is held at zero, and*
(d) $O(n^{8/3}k^{16/3})$ *under the similarity measure if the gap penalty is allowed to vary.*

Proof. By Lemma 5 and the fact that the number of internal nodes of any phylogeny for $\mathcal{S}$ is at most k, $0 \leqslant w_{\mathcal{A}}, x_{\mathcal{A}}, y_{\mathcal{A}}, z_{\mathcal{A}} \leqslant N = O(nk^2)$. Under distance measures, the total score of an alignment is $x_{\mathcal{A}} + \beta y_{\mathcal{A}} + \gamma z_{\mathcal{A}}$. Now, (a) and (b) follow from Lemma 8 with $d = 1$ and $d = 2$, respectively. Under similarity measures, the score is $w_{\mathcal{A}} - \alpha x_{\mathcal{A}} - \beta y_{\mathcal{A}} - \gamma z_{\mathcal{A}}$. Thus, parts (c) and (d) follow from Lemma 8 with $d = 2$ and $d = 3$, respectively. □

The bounds for phylogenetic alignment can be improved if the input phylogeny is a star.

Theorem 17. *The number of optimality regions for star alignment under global alphabet-independent scoring is $O(n^{2/3}k^{2/3})$ when the gap penalty is zero and $O(n^{5/3}k^{5/3})$ when the gap penalty varies.*

Proof. By Lemma 5 with $r = 1$, $0 \leqslant w_{\mathcal{A}}, x_{\mathcal{A}} \leqslant nk$, while by Lemma 6, $y_{\mathcal{A}}, z_{\mathcal{A}} \leqslant nk$. The bounds now follow from Lemma 8 with $N = nk$, and $d = 2$ and $d = 3$, respectively. □

Finally, we prove polynomial bounds for the alphabet-dependent case when the alphabet size is fixed. As in Theorem 16, our bounds are the same for phylogenetic and generalized phylogenetic alignment.

Theorem 18. *The number of optimality regions for optimality regions for alphabet-dependent parametric phylogenetic and generalized phylogenetic alignment under distance and similarity measures is* $(nk^2)^{O(|\Sigma|^2)}$.

Proof. By Eq. (5), the similarity score of an alignment $\mathcal{A}$ is

$$\sigma_{\mathcal{A}} = \gamma \sum_{(u,v)\in T} z_{uv} + \sum_{\{s,t\}\subseteq\Sigma} \alpha(s,t) \sum_{(u,v)\in T} x_{uv}(s,t), \tag{14}$$

where T is the phylogeny, z_{uv} and $x_{uv}(s,t)$ are, respectively, the number of gaps and the number of times character s is lined up with character t in the pairwise alignment between the strings labeling nodes u and v of T. The value of $\mathcal{A}$ is thus a function of $O(|\Sigma|^2)$ parameters. Moreover, $\sum x_{uv}(s,t)$ can take on $O(nk^2)$ distinct values. The claim now follows from Lemma 9. □

5. Discussion

Since the number of optimality regions for all problems considered here is polynomial in the length and number of sequences (assuming bounded alphabet in the alphabet-dependent case), Theorem 10 implies that the corresponding minimization diagrams can be computed with a polynomial number of calls to algorithms for the respective fixed-parameter problems. However, the fact that an exact solution to these problems is needed is a severe limitation, since the cost of carrying out even a single multiple or phylogenetic alignment is prohibitive, except for short sequences.

In practice, the fixed-parameter problems are often solved heuristically, and each such scheme raises its own parameter-sensitivity issues. The approach presented here can be used to analyze any scheme that minimizes or maximizes a function that depends on discrete features of pairwise alignments. Examples of such approaches are given in [18]. Different techniques seem necessary to analyze heuristics that do not fall in this category; e.g., progressive alignment [12] and some of the methods outlined in [35].

Acknowledgements

Thanks to Gavin Naylor for discussions on sensitivity analysis in phylogeny construction and for pointing out [38]. Thanks also to Steven Altschul for clarifying the meaning of gaps in multiple alignments and to Dan Gusfield for hospitality and useful discussions while the first author visited Davis.

References

[1] P.K. Agarwal, M. Sharir, Davenport–Schinzel Sequences and their Geometric Applications, Cambridge University Press, Cambridge, 1995.
[2] S.F. Altschul, Gap costs for multiple sequence alignment, J. Theor. Biol. 138 (1989) 297–309.
[3] S.F. Altschul, Leaf pairs and tree dissections, SIAM J. Discrete Math. 2 (3) (1989) 293–299.
[4] S.F. Altschul, R.J. Carrol, D.J. Lipman, Weights for data related by a tree, J. Mol. Biol. 207 (1989) 647–653.
[5] S.F. Altschul, D.J. Lipman, Trees, stars, and multiple biological sequence alignment, SIAM J. Appl. Math. 49 (1) (1989) 197–209.
[6] M.O. Dayhoff, R.M. Schwartz, B.C. Orcutt, A model of evolutionary change in proteins, Atlas of Protein Sequence and Structure 5 (1978) 345–352.
[7] D. Dobkin, H. Edelsbrunner, C.K. Yap, Probing convex polytopes, in: Proceedings of the 18th Annual ACM Symposium on Theory of Computing, 1986, pp. 424–432.
[8] D. Dobkin, H. Edelsbrunner, C.K. Yap, Probing convex polytopes, in: Cox, Wilfong (Eds.), Autonomous Robot Vehicles, Springer-Verlag, 1990, pp. 328–341.
[9] H. Edelsbrunner, Algorithms in Combinatorial Geometry, Springer-Verlag, Heidelberg, 1987.
[10] M.J. Eisner, D.G. Severance, Mathematical techniques for efficient record segmentation in large shared databases, J. Assoc. Comput. Mach. 23 (1976) 619–635.
[11] J. Felsenstein, Phylogeny programs, http://evolution.genetics.washington.edu/phylip/software.html.
[12] D. Feng, R.F. Doolittle, Progressive alignment as a prerequisite to correct phylogenetic trees, J. Mol. Evol. 25 (1987) 351–360.
[13] D. Fernández-Baca, T. Seppäläinen, G. Slutzki, Bounds for parametric sequence comparison, in: Proc. Symp. on String Processing and Information Retrieval, IEEE Computer Society, 1999, pp. 55–62.
[14] D. Fernández-Baca, S. Srinivasan, Constructing the minimization diagram of a two-parameter problem, Oper. Res. Lett. 10 (1991) 87–93.
[15] S.K. Gupta, J. Kececioglu, A.A. Schäffer, Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment, J. Comput. Biol. 2 (1995) 459–472.
[16] D. Gusfield, Sensitivity analysis for combinatorial optimization, Technical Report UCB/ERL M80/22, University of California, Berkeley, May 1980.
[17] D. Gusfield, Parametric combinatorial computing and a problem in program module allocation, J. Assoc. Comput. Mach. 30 (3) (1983) 551–563.
[18] D. Gusfield, Efficient algorithms for multiple sequence alignment with guaranteed error bounds, Bull. Math. Biol. 55 (1993) 141–154.
[19] D. Gusfield, Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology, Cambridge University Press, Cambridge, 1997.
[20] D. Gusfield, K. Balasubramanian, D. Naor, Parametric optimization of sequence alignment, Algorithmica 12 (1994) 312–326.
[21] D. Gusfield, R.W. Irving, Parametric stable marriage and minimum cuts, Inform. Process. Lett. 30 (1989) 255–259.
[22] D. Gusfield, P. Stelling, Parametric and inverse-parametric sequence alignment with XPARAL, in: R.F. Doolittle (Ed.), Computer methods for macromolecular sequence analysis, in: Methods in Enzymology, vol. 266, Academic Press, 1996, pp. 481–494.
[23] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. USA 89 (1992) 10915–10919.
[24] X. Huang, P.A. Pevzner, W. Miller, Parametric recomputing in alignment graphs, in: M. Crochemore, D. Gusfield (Eds.), Combinatorial Pattern Matching, in: Lecture Notes in Computer Science, vol. 807, Springer-Verlag, 1994, pp. 87–101.
[25] T. Jiang, L. Wang, On the complexity of multiple sequence alignment, J. Comput. Biol. 1 (1994) 337–348.
[26] T. Jiang, E.L. Lawler, L. Wang, Aligning sequences via an evolutionary tree: Complexity and approximation (extended abstract), in: Proc. 26th Ann. Symp. Theory of Computing, 1994, pp. 760–769.
[27] W. Just, Computational complexity of multiple sequence alignment with SP-score, unpublished.
[28] W. Just, G. Della Vedova, Multiple sequence alignment as a facility location problem, in: Proceedings of the Prague Stringology Club Workshop, 2000.

[29] D.J. Lipman, S.F. Altschul, J.D. Kececioglu, A tool for multiple sequence alignment, Proc. Natl. Acad. Sci. USA 86 (1989) 4412–4415.
[30] F.P. Preparata, M.I. Shamos, Computational Geometry: An Introduction, Springer-Verlag, 1985.
[31] D. Sankoff and R. J. Cedergren, Simultaneous comparison of three or more sequences related by a tree, in [33], pp. 253–263.
[32] D. Sankoff, Minimal mutation trees of sequences, SIAM J. Appl. Math. 28 (1) (1975) 35–42.
[33] D. Sankoff, J.B. Kruskal (Eds.), Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, Addison-Wesley, Reading, MA, 1983.
[34] T. Shibuya, H. Imai, New flexible approaches for multiple sequence alignment, J. Comput. Biol. 4 (3) (1997) 385–413.
[35] M. Vingron, Sequence alignment and phylogeny construction, in: M. Farach-Colton, et al. (Eds.), Mathematical Support for Molecular Biology, in: DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 47, American Mathematical Society, 1998.
[36] M. Vingron, M.S. Waterman, Sequence alignment and penalty choice: Review of concepts, case studies, and implications, J. Mol. Biol. 235 (1994) 1–12.
[37] M.S. Waterman, M. Eggert, E. Lander, Parametric sequence comparisons, Proc. Natl. Acad. Sci. USA 89 (1992) 6090–6093.
[38] W.C. Wheeler, Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data, Syst. Bio. 44 (3) (1995) 321–331.